Published in final edited form as:
Neuroinformatics. 2018 October ; 16(3-4): 285–294. doi:10.1007/s12021-018-9381-1.

Cognitive Assessment Prediction in Alzheimer's Disease by Multi-Layer Multi-Target Regression

Xiaoqian Wang[1], **Xiantong Zhen**[1], **Quanzheng Li**[2], **Dinggang Shen**[3], and **Heng Huang**[1]

[1]Department of Electrical, Computer Engineering, University of Pittsburgh, Pennsylvania, PA 15263, USA

[2]Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

[3]Radiology and BRIC, UNC-CH School of Medicine, 130 Mason Farm Road, Chapel Hill, NC 27599, USA

Abstract

Accurate and automatic prediction of cognitive assessment from multiple neuroimaging biomarkers is crucial for early detection of Alzheimer's disease. The major challenges arise from the nonlinear relationship between biomarkers and assessment scores and the inter-correlation among them, which have not yet been well addressed. In this paper, we propose multi-layer multi-target regression (MMR) which enables simultaneously modeling intrinsic inter-target correlations and nonlinear input-output relationships in a general compositional framework. Specifically, by kernelized dictionary learning, the MMR can effectively handle highly nonlinear relationship between biomarkers and assessment scores; by robust low-rank linear learning via matrix elastic nets, the MMR can explicitly encode inter-correlations among multiple assessment scores; moreover, the MMR is flexibly and allows to work with non-smooth $l_{2,1}$-norm loss function, which enables calibration of multiple targets with disparate noise levels for more robust parameter estimation. The MMR can be efficiently solved by an alternating optimization algorithm via gradient descent with guaranteed convergence. The MMR has been evaluated by extensive experiments on the ADNI database with MRI data, and produced high accuracy surpassing previous regression models, which demonstrates its great effectiveness as a new multi-target regression model for clinical multivariate prediction.

Keywords

Multi-target regression; Robust low-rank learning; Calibration; Nonlinear regression; Alzheimer's disease

Heng Huang henghuanghh@gmail.com. Xiaoqian Wang xqwang1991@gmail.com. Xiantong Zhen zhenxt@gmail.com. Quanzheng Li li.quanzheng@mgh.harvard.edu. Dinggang Shen dgshen@med.unc.edu.

Introduction

Alzheimer's disease (AD) is the most common cause of dementia and is characterized by progressive loss of memory. AD severely impacts human thinking and behavior. The influence of AD is both extensive and complex, making it difficulties to prevent or diagnose this disease (Association et al. 2016). Neuroimaging techniques provide a powerful tool for the early diagnosis and response monitoring of Alzheimer's such that the diagnostic capabilities can be improved. The Alzheimer's Disease Neuroimaging Initiative (ADNI) study (Jack et al. 2008; Mueller et al. 2005) collects neuroimaging and cognitive measurement of normal aging, mild cognitive impairment as well as AD samples, which provides a wealth of resources for the study of Alzheimer's diagnosis, treatment and prevention.

According to the statistics in Association et al. (2016), by 2016 Alzheimer's affects a total of 5.4 million American people. This disease is incurable and there is no accurate way to diagnose AD. Thus, the current consensus emphasizes the need to diagnose and explore cognitive performance of brain function. Several cognitive tests have been presented to assess individual's cognitive level, such as Mini-Mental State Examination (MMSE) (Folstein et al. 1975) and Rey Auditory Verbal Learning Test (RAVLT) (Schmidt and et al. 1996). In recent AD research, a wide range of work has employed regression models to uncover the relationship between neuroimaging data and cognitive test scores (Seshadri et al. 2007; Moradi et al. 2016). The major challenges arise from the jointly modeling nonlinear relationship between biomarkers and assessment scores and the intrinsic correlation among assessment scores.

Since the biomarkers are features extracted from imaging data and therefore contain relatively low-level information, while the assessment scores are high-level measurement of disease progress, the relationship between biomarkers and scores tends be complex and highly nonlinear. However, most previous methods use linear regression models to predict the relationship between imaging biomarkers and cognitive assessment (Ferrarini et al. 2008; Moradi et al. 2016), which is not appropriate to illustrate the complex influence of brain structure impairment.

Since different cognitive assessment scores provide the measurement related to the same disease progress, these scores as regression targets are correlated. Finding out such inter-target relations can be beneficial to analyzing the influence of neuroimaging biomarkers on memory assesment performance. In this paper, we address cognition assessment by formulating it as a multi-task learning problem with a newly proposed multi-target regression (*a.k.a.,* multi-output regression) model. Although multi-target regression has been extensively explored, existing models have some shortcomings and would not achieve satisfactory performance on our specific application.

To explore inter-target correlations, existing multi-target regression models were focused mainly on linear regression models (Rothman et al. 2010; Sohn and Kim 2012; Rai et al. 2012; Gong et al. 2014; Liu et al. 2014; Pan et al. 2015; Zhu et al. 2017) or specifically developed under particular assumptions with prior knowledge (Argyriou et al. 2008;

Agarwal et al. 2010; Zhang and Yeung 2013; Kumar and Daume 2012; Ciliberto et al. 2015), which facilitates the correlation modeling. By building upon linear regression models, sparsity or low rank is simply imposed on the regression matrix to capture inter-target correlations (Kolar et al. 2011; Liu et al. 2014; Molstad and Rothman 2015); however, these linear models suffer from the limited ability to handle nonlinear relationships between high-dimensional inputs and multiple targets (Hara and Chellappa 2014), and moreover it is non-trivial to extend these linear models for nonlinear regression due to the non-convexity of sparsity constraints or loss functions (Liu et al. 2014; Dinuzzo and Schölkopf 2012). By making some specific assumptions, e.g., regression task parameters share a common prior (Yu et al. 2005; Lee et al. 2007; Daume´ III 2009), or lie in on a low-dimensional manifold (Agarwal et al. 2010) or share a linear subspace (Kumar and Daume 2012), particular inter-target correlations were explored in previous work (Yu et al. 2005; Lee et al. 2007; Daume´ III 2009; Agarwal et al. 2010; Kumar and Daume 2012); however, these assumptions can be too restrictive and would not necessarily hold or be shared by different applications in practice (Zhang and Yeung 2014), which makes them lack of generality.

To handle the complex nonlinear input-output relationships, kernel methods (Evgeniou et al. 2005; Alvarez et al. 2012; Li et al. 2015) were extended from single task learning to multi-task learning. In Evgeniou et al. (2005), the regression matrix of multiple tasks is simply reshaped into a vector to explore inter-target correlations, which however does not distinguish between inter and intra tasks and tends to be less effective to encode the correlations. Moreover, the method assumes that the task similarity between tasks is given and the regularization term is based on the similarity which however is mostly unknown and varies dramatically with different applications. In addition, since the similarity is nonnegative, the model can only model positive relationships between multiple tasks.

To achieve this goal, in this paper we propose a novel model, Multi-layer Multi-target Regression (MMR). Our model enables simultaneously modeling intrinsic inter-target correlations and complex input-output relationships in one single general framework. The MMR accomplishes a multi-layer learning architecture which is composed of the input, hidden and target (output) layers as illustrated in Fig. 1.

The proposed MMR leverages the strength of kernel methods for nonlinear feature learning and the structural advantage of multi-layer architectures to capture inter-target correlations, which could explicitly encode the correlations among different cognitive learning tasks. More importantly, it provides a new multi-layer learning paradigm that is endowed with high generality, flexibility and expressive ability for multi-target clinical data prediction.

The contributions of this work are summarized as follows:

- We formulate cognitive assessment as a multi-task learning problem, which is fulfilled by a newly proposed multi-layer multi-target regression (MMR) model.
- We introduce the compositional learning framework which enables jointly modeling nonlinear input-output relationship and intrinsic inter-target correlations.

- We introduce the $l_{2,1}$-norm loss function to achieve automatic calibration of multiple targets with disparate noise levels, which enables more robust parameter estimation.

Related Work

In this section, we briefly review related work in terms of both multi-target regression and the Alzheimer's disease application.

Since AD is a chronic neuro-degenerative disease, it is important to reveal the correlation between changes in brain structure and cognitive dysfunction. Recent studies have employed different machine learning models to analyze the association between imaging markers and cognitive performance. In Ferrarini et al. (2008), the authors employed linear regression models to evaluate the correlation between structural brain atrophy and MMSE cognitive score, where well-defined periventricular structures like left temporal horn, left corona radiata, and the right caudate nuclei were found to have distinct impact on the performance in MMSE cognitive test.

Moradi et al. (2016) applied elastic net linear regression in the prediction of RAVLT test score via MRI data. They identified several neuroimaging features for the estimation of RAVLT, including medial temporal lobe structures angular gyrus, hippocampus and amygdala, which shed insights on understanding the influence of these important brain regions for episodic memory.

Falahati et al. (2016) conducted a longitudinal investigation among an non-demented and stroke-free people from the Rotterdam study. By means of linear and Cox regression models, the authors revealed the correlation between hippocampal subiculum and the onset of dementia, which indicated an important marker for dementia prediction.

Zhu et al. (2015) combined support vector regression (SVR) and support vector machine (SVM) to jointly predict the clinical scores as well as the disease status. They formulated the joint learning process in a multi-task learning framework such that different tasks will strengthen each other in AD diagnosis.

The successful applications of machine learning approaches in the prediction of cognitive impairment strengthened the study of underlying pathology in Alzheimer's. In a cognitive assessment, there are usually several different tests involved. The output from different tests can be correlated. Rothman et al. (2010) put forward an approach to explore the output structure. They proposed a multivariate regression model with covariance estimation (MRCE), in which a procedure is developed for constructing a sparse estimator of a multivariate regression coefficient matrix that accounts for correlation of the response variables. However, the MRCE does not leverage the learned output structure to share similar input variables among related outputs (Sohn and Kim 2012). Moreover, it is a linear regression model with limited ability to handle nonlinear regression tasks.

Moreover, if we treat the estimation of each cognitive test score as one task, we can naturally formulate the situation of multiple cognitive test estimation as a multi-task learning problem.

Previous machine learning models provided different approaches on how to automatically capture the structure among tasks. In Zhang and Yeung (2014), Zhang and Yeung proposed a convex formulation for multi-task relationship learning (MTRL), which models the relationships between tasks in a nonparametric manner based on the assumption that all tasks are close to each other by measuring the Frobenius norms of their differences. The MTRL is developed based on prior assumptions of multivariate normal distributions on both multiple targets and regression parameters. However, those assumptions do not necessarily hold in practice or be shared by different applications.

The MTRL is further generalized in Rai et al. (2012) where multi-target regression with output and task structures (MROTS) is proposed to jointly explore the covariance structure of latent model parameters and the conditional covariance structure of multiple targets. MROTS outperforms both MTRL and MRCE, which however similar to the MRCE (Rothman et al. 2010) does not admit trivial extensions to nonlinear regression. Sohn and Kim (2012) introduce a matrix l_1 norm based inverse-covariance regularization for joint estimation of structured sparsity and output structure for multi-target regression, where the output structure of multiple targets is represented as a graph.

Inter-target correlation has also been investigated in kernel scenarios. the output kernel learning (OKL) was developed for vector-valued functions to explore inter-target correlations for multiple task learning (Alvarez et al. 2012; Dinuzzo et al. 2011; Dinuzzo 2013). Nevertheless, the OKL does not fully capture inter-target correlations since it simply learns a semi-definite similarity matrix of multiple targets.

Multi-target regression has also been studied under the framework of ensemble learning. A fitted rule ensembles (FIRE) algorithm is introduced in Aho et al. (2012) to improve multi-target regression by adding simple linear functions to the ensemble. Based on ensemble learning, Tsoumakas et al. (2014) construct new target variables by random linear combination (RLC) of existing targets, which is heuristically derived from multi-label classification. However, those methods fail to take into account the correlation of multiple targets.

Recently, Zhou and Zhao (2016) propose flexible clustered multi-task (FCMTL) which is an improved version of clustered multi-task learning (CMTL). In order to explore inter-target correlation, also based on the cluster assumption, the cluster structure is learned in FCMTL by identifying representative tasks. However, the assumption of the existence of representative tasks would be too strong and not necessarily shared by different applications due to the diversity.

Multi-Layer Multi-Target Regression

Multi-Target regression is to learn a holistic mapping function h from the input space $X \in \mathbb{R}^d$ to the multivariate target (output) space $y \in \mathbb{R}Q$, where d is the dimensionality of the input space and Q is the number of targets. We will find a function h that is able to simultaneously handle the aforementioned multiple challenges within one single framework by a general compact formulation of compositional learning.

A General Compositional Learning Framework

Given a training set of training data $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N\}$ associated with targets $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_N\}$, the MMR seeks a mapping function h: $X \rightarrow Y$, which takes a generic formula as follows

$$\mathbf{y} = h(\mathbf{x}) = f(g(\mathbf{x})) = (f \; o \; g)(\mathbf{x}) \quad (1)$$

where g and f are two functions that are jointly learned to establish the overall mapping h from input image representations to target shapes. g serves to extract high-level features, i.e., the latent variables, $\mathbf{z}$ that span the latent space, where $\mathbf{z} \in \mathbb{R}^Q$,

$$\mathbf{z} = g(\mathbf{x}) \quad (2)$$

From the latent space, we propose explicitly encoding the inter-target correlations via f by:

$$\mathbf{y} = f(\mathbf{z}) \quad (3)$$

Both g and f can be customized according to diverse applications and regulated by different constraints to favor specific properties, which ensures the generality of the MMR. It establishes a multi-layer learning architecture which is endowed with high generality, great flexibility and strong expressive ability to jointly handle highly complex input-output relationships and intrinsic correlations of multiple targets in one single framework.

The functions f and g can be found by the following general regularized learning framework:

$$\min_{f, g \in \mathscr{H}K} \frac{1}{\mathcal{N}} \sum_{i}^{\mathcal{N}} \mathscr{L}\left(f, g, \mathbf{x}_i, \mathbf{y}_i\right) + \lambda \Omega(g) + \beta \Omega(f) \quad (4)$$

where $\mathscr{L}$ is the general loss function which could be the least square error or the hinge loss; λ and β are the regularization parameters; $\Omega(g)$ is the regularization term g to control its complexity to prevent overfitting; $\Omega(f)$ is the regularization term on S to encode intrinsic inter-target correlations. The latent variables can be viewed as higher-level features that facilitate jointly modeling input-output relationships and inter-target correlation. In the following, we specify functions f and g to achieve the multi-layer multi-target regression (MMR).

Nonlinear Learning via Kernelized Dictionary

We propose building the nonlinear function g via Kernelized dictionary rather than based on the kernel extension of a linear regression model (Zhen et al. 2017). Specifically, the function g takes the following forms:

$$\mathbf{z} = \sum_{i=1}^{N} \alpha i K\left(\mathbf{x}_i, \mathbf{x}\right) = AK_{\mathbf{x}} \quad (5)$$

Where $K\left(\mathbf{x}_i, \cdot\right)_{i=1}^{N}$ is the Kernelized dictionary, $k_{\mathbf{x}} = [\ldots; k(\mathbf{x}i, \mathbf{x}); \ldots] \in \mathbb{R}^N$; $\boldsymbol{\alpha} \in \mathbb{R}^Q$ is the coefficients associated with each atom in the dictionary and $A \in \mathbb{R}^{Q \times N} = [\boldsymbol{\alpha}1, \ldots, \boldsymbol{\alpha}i, \ldots, \boldsymbol{\alpha}_N]$. To achieve nonlinear learning, we usually employ the radius basis function as the kernel function, i.e., $k(\mathbf{x}i, \mathbf{x}j) = \exp(\|\mathbf{x}i - \mathbf{x}j\|^2/\sigma^2)$, where σ is the band width.

It is notable that (5) is the multivariate extension of the conventional kernel dictionary (Feng et al. 2016), which has recently drawn great attention. The advantages of using kernelized dictionary for nonlinear learning rather than kernel extension (Zhen et al. 2017) lie in two major aspects. On one hand, the kernel k is not necessarily the Mercer kernel, which makes learning more applicable because the Mercer condition on the kernel may be difficult to satisfy. On the other hand, it allows to more flexibly design loss functions, not necessarily restricted to strictly smooth functions. This benefit will be shown in the calibration of multiple targets (Section 3).

If we choose the Frobenius norm loss, the objective function (4) turns out to be the form as follows

$$\min_{f, A} \frac{1}{\mathcal{N}} \left\| Y - AK \right\|_F^2 + \lambda \left\| A \right\|_F^2 + \beta \Omega(f) \quad (6)$$

The multi-target regression model in (6) is decoupled into several single-target problems, which does not take into account inter-target correlations, resulting in suboptimal multi-target regression with inferior performance. In what follows, we introduce our MMR, which is a multi-layer learning architecture to explicitly model the correlations by a robust low-rank learning with matrix elastic nets (MEN).

Robust Low-Rank Learning via Matrix Elastic Nets

Rather than directly imposing sparsity regression coefficients in existing methods, we propose incorporating a structure matrix S to explicitly encode inter-target correlations via a rank minimization.

$$\min_{W, S} \frac{1}{\mathcal{N}} \left\| Y - SZ \right\|_F^2 + \lambda \left\| A \right\|_F^2 + \beta Rank(S) + \gamma \left\| S \right\|_F^2, \quad (7)$$

where $Z = [\mathbf{z}_1, \ldots, \mathbf{z}_i, \ldots, \mathbf{z}^N] \in \mathbb{R}^{Q \times N}$, $\mathbf{z}_i = Ak_{\mathbf{x}i} \in \mathbb{R}_Q$ contains the latent variables in the latent space, $S \in \mathbb{R}^{Q \times Q}$ is the structure matrix that serves to explicitly model inter-target correlations, β is the regularization parameter to control the rank of S, that is, a larger β induces lower rank, and the Frobenius norm control the shrinkage of S with the associated parameter γ. The rank minimization of the structure matrix *S* explores the low-rank structure existing between tasks to capture the intrinsic inter-target correlation. *S* is learned automatically from data without relying on any specific assumptions, which allows to adaptively cater different applications.

However, the objective function in (7) is NP-hard due to the noncontinuous and non-convex nature of the rank function. The nuclear norm $\|S\|_*$ is commonly used and has been proven to be the convex envelop of the rank function over the domain $\|S\|_2 \leq 1$, which provides the tightest lower bound among all convex lower bounds of the rank function Rank*(S)*.

As a consequence, the combination of the nuclear norm with the Frobenius norm on *S* gives rise to the matrix elastic net (MEN) (Li et al. 2012) as a regularizer of (7):

$$\min_{W,S} \frac{1}{\mathcal{N}} \|Y - SZ\|_F^2 + \lambda \|A\|_F^2 + \beta \|S\|_* + \gamma \|S\|_F^2, \quad (8)$$

where the nuclear norm $\|S\|_*$ is also known as the trace norm.

The MEN is an analog to the elastic-net regularization (Zou and Hastie 2005) from compressive sensing and sparse representation (Zou and Hastie 2005). It has been shown that the elastic net often outperforms the lasso (Zou and Hastie 2005). In the MEN, the nuclear-norm constraint enforces the low-rank property of the solution *S* to encode inter-target correlations, and the Frobenius-norm constraint induces a linear shrinkage on the matrix entries leading to stable solutions (Li et al. 2012). The MEN regularization generalizes the matrix lasso and provides improved performance than lasso (Tibshirani 1996). To the best of our knowledge, this is the first work that introduces the MEN to multi-target regression for robust low-rank learning, which offers a general framework to encode inter-target correlations.

Calibration of Multiple Targets

Multivariate targets (outputs) exhibit distinct noise levels which are usually unknown *a priori* in practice (Rakitsch et al. 2013; Gillberg et al. 2016). It is theoretically shown that regularization parameters should be chosen in proportion to the maximum standard deviations of the noise for each target to achieve optimal parameter estimation error bound (Lounici et al. 2011). It is crucial to take into account the disparate noise levels of multivariate targets to achieve robust parameter estimation for improved prediction performance.

We replace the Frobenius norm in (8) with the $l_{2,1}$-norm as the loss function to calibrate multivariate targets (Liu et al. 2014), which accomplishes the final objective function as follows:

$$\min_{A,S} \frac{1}{\mathcal{N}} \|Y - SAK\|_{2,1} + \lambda tr\left(A^\top A\right) + \beta tr\left(\sqrt{S^\top S}\right) + \gamma tr\left(S^\top S\right). \quad (9)$$

In (9), the induced latent variables $Z = Ak$ can extract high-level representations for multiple semantic targets, which allows to disentangle the nonlinear relationship between low-level inputs and semantic-level targets. The latent variables are represented by notes of hidden layers in Fig 1. The latent space with high-level features will also facilitate the efficient linear low-rank learning of S to model inter-target correlations to achieve more accurate multi-target prediction. The MMR in (9) leverages the strength of kernel methods for nonlinear feature extraction and the structural advantage of multi-layer architectures for inter-target correlation modeling. In contrast to existing multi-target regression models, the obtained MMR in (9) accomplishes a new multi-layer learning architecture, which 2 min is endowed with great generality, flexibility and expressive ability for diverse challenging tasks. We derive a new alternating optimization algorithm to efficiently solve the objective in (9), which associated with the convergence proof is attached in the supplementary material due to the space limit.

Alternating Optimization

The obtained objective function (9) is non-trivial to solve simultaneously for A and S due to the non-convexity of the objective function. We derive a new alternating optimization algorithm to efficiently solve the objective function. Denote $J(A, S)$ as the objective function in (9), and we seek A and S alternately by solving $J(A, S)$ for one with the other fixed.

Fix S to Optimize A

We calculate the gradients of the objective function with respect to A as follows:

$$\frac{\partial J}{\partial A} = -\frac{1}{\mathcal{N}}S^{\top}G(\Delta)K + \lambda A\,. \quad (10)$$

Where

$$(G)_{ii} = \frac{1}{2\|\Delta\|_2} \quad (11)$$

and $\Delta = Y - SAk$. Denote $\mathcal{G}(A)\frac{\partial J}{\partial A}$. A is updated by gradient descent as

$$A^{t+1} = A^{t} - \eta_t \mathcal{G}\left(A^{t}\right). \quad (12)$$

where η_t is the learning rate which can adaptively chosen by line search algorithms (Armijo 1966).

Fix A to Optimize S

We propose a gradient based alterative optimization to solve for *S*, before which we provide the following proposition to calculate the derivative of *J* w.r.t. *S*.

Proposition 1 *Assume that the singular value decomposition (SVD) of S is*

$$S = U \sum V^\top, \quad (13)$$

where U and V are unitary matrices and Σ is the diagonal matrix with real numbers on the diagonal. Then the derivative of $\|S\|_*$ *w.r.t. S takes the form as follows:*

$$\frac{\partial \|S\|_*}{\partial S} = UV^\top \quad (14)$$

The proof is provided in the Appendix section.

Proposition 1 associated with the rigorous proof provides a theoretical foundation, which can be directly used to solve a large while important family of optimization problems with trace norm minimization.

Based on the Proposition 1, we have the derivative of *J* w.r.t *S* as follows:

$$\frac{\partial J}{\partial S} = -2\frac{1}{\mathcal{N}}(Y - SAK)(AK)^\top + \beta UV^\top + 2\gamma S \quad (15)$$

where *U* and *V* are obtained by the SVD in (13).

Denote *G(S)* as the gradient w.r.t. *S* in (15). Therefore, *S* can be solved by an iterative optimization based on gradient descent.

$$S^{t+1} = S^t - \eta_t \mathcal{G}\left(S^t\right) \quad (16)$$

where η_t is the learning rate, which can be adaptively chosen by line search algorithms (Armijo 1966). In each iteration, S^{t+1} is calculated with the current S^t associated with *U*, *Σ* and *V*. Since the objective function *J (A, S)* is convex with respect to *S*, it is guaranteed to find a global minimum of *S*.

Note that the size of *S* depends only on the number *2* of targets, which is usually much smaller than the dimensionality *d* of inputs. Therefore, the complexity of the singular value decomposition (SVD) of the structure matrix *S* involved in the calculation of the derivative of the nuclear norm is O *(Q^3)*. This guarantees the efficiency of both the iterative algorithm to update *S* and the alternating optimization algorithm (Algorithm 1).

```
Algorithm 1 Alternating Optimization
Require: Data matrices $X$ associated with corresponding
    targets $Y$, regularization parameters $\lambda$, $\beta$ and and $\gamma$.
Ensure: The regression coefficient matrix $A$ and the
    structure matrix $S$.
Randomly initialize $S \in \mathbb{R}^{Q \times Q}$ and set $i = 1$;
repeat
    Update A using (12);
    Update S using (16);
    $i \leftarrow i + 1$;
until Convergence.
```

Convergence Analysis

The efficiency of the proposed MMR is ensured by the guaranteed convergence of the newly-derived alternating optimization algorithm. The objective function $J(A, S)$ in Section 3 is bounded from below and monotonically decreases with each optimization step for A and S, and therefore it converges. We give the brief sketch of the convergence analysis.

Since $J(A, S)$ is the summation of norms, we have $J(A, S) \geq 0$ for any A and S. Then $J(A, S)$ is bounded from below. Denote $A^{(t)}$ and $S^{(t)}$ as the A and S in the t-th iteration, respectively. For the t-th step, $A^{(t)}$ is computed by $A^{(t)} \leftarrow \arg\min_A J(A, S^{(t-1)})$. And we also have $J(A^{(t)}, S^{(t-1)}) \geq J(A^{(t)}, S^{(t)})$. In this way, we obtain the following inequality:

$$\cdots \geq J(A^{(t-1)}, S^{(t-1)}) \geq J(A^{(t)}, S^{(t-1)}) \geq J(A^{(t)}, S^{(t)}) \geq \cdots .$$

Therefore, $J(A^{(t)}, S^{(t)})$ is monotonically decreasing as $t \rightarrow +\infty$, which indicates that the objective function $J(A, S)$ converges according to the monotone convergence theorem.

Experiments and Results

In this section, we conduct extensive experiments to test the performance of the proposed multi-layer multi-target regression (MMR) model in predicting cognitive scores on public ADNI data. We provide comprehensive comparison with representative multi-output regression models to show the advantages of the MMR. The experimental results have demonstrated the effectiveness of the MMR for cognitive assessment of AD.

Data Description

The data used in this article comes from the ADNI database (adni.loni.usc.edu). Firstly, for each MRI T1-weighted image, we corrected the anterior commissure (AC) posterior commissure (PC) via MIPAV2; corrected the intensity inhomogeneity using N3 algorithm (Sled et al. 1998); stripped the skull (Wang et al. 2011) with manual editing, and removed the cerebellum (Wang et al. 2014). Afterwards, we divided the image into gray matter (GM), white matter (WM), as well as cerebrospinal fluid (CSF) by means of FAST (Zhang et al. 2001) in the FSL package3, and then used HAMMER (Shen and Davatzikos 2002) to register the images to a common space. The GM volumes that were obtained from 93 ROIs defined in Kabani (1998), normalized by the total intra-cranial volume, were characterized as features. We downloaded the cognitive scores from three independent cognitive assessments, including Fluency Test, Rey's Auditory Verbal Learning Test (RAVLT) and Trail making test (TRAILS). We suggest interested readers to find the details of these cognitive assessments in the ADNI procedure manuals. All participants with no missing baseline MRI measurements and cognitive measures were included in this study. A total of 804 sample subjects were considered, of which we have 225 health control (HC) samples, 393 MCI samples and 186 AD samples. This study involved seven cognitive scores, which are: 1) RAVLT TOTAL, RAVLT TOT6 and RAVLT RECOG scores from RAVLT cognitive assessment; 2) FLU ANIM and FLU VEG scores from Fluency cognitive assessment; 3) Trails A and Trails B scores from Trail making test.

Experimental Settings

To evaluate the performance of our model, **MMR** (multi-layer multi-target regression), we compare with several representative regression models as used in Wang et al. (2012), which includes least square regression (**LSR**), multi-target ridge regression (**MRR**) and multi-target low-rank regression model with trace norm regularization (**MR-Trace**).

To make contrast to our multi-layer multi-target regression (MMR), we give the formulation of the baseline methods. The LSR takes the following form of objective,

$$\min_{W} \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \left\| \mathbf{y}_i - W\mathbf{x}_i \right\|^2, \quad (17)$$

where $W \in \mathbb{R}^{Q \times d}$ is the weight matrix;

MMR is the baseline kernel method which takes the following form

$$\min_{W} \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \left\| \mathbf{y}_i - W\mathbf{x}_i \right\|^2 + \lambda \left\| W \right\|_F^2, \quad (18)$$

where $\left\| W \right\|_F^2$, is the regularization term to avoid overfitting and λ is the hyper-parameter for the regularization term;

The MR-Trace is a single-layer learning model with a trace norm regularization:

$$\min_{W} \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \left\| \mathbf{y}_i - W\mathbf{x}_i \right\|^2 + \lambda \left\| W \right\|_*, \quad (19)$$

where the trace norm $\| W \|_*$ is the trace norm regularization term to impose sparsity constraint.

The major methodological differences between the baseline and the proposed method is our multi-layer learning architecture. Specifically, our MMR incorporates a hidden layer of latent variables rather than directly projecting the input data to the multiple outputs as in the baseline methods. This actually enables us to simultaneously handle nonlinear input-output relationship (thorough the kernelized dictionary) and the interdependency of multiple outputs (by low-rank learning). The major methodological differences indeed explains the improvement of our method over baseline methods.

In the experiment, we use the root mean square error (RMSE) and the correlation coefficient (CorCoe) between the predicted value and ground truth as the evaluation measurements. We normalize the RMSE value with the Frobenius norm of the ground truth. To illustrate the stability of the comparing methods, we adopt 5-fold cross validation and report average performance in these 5 trials.

For MMR model, we choose the regularization parameters by cross validation. We tune other hyper-parameters, i.e., parameter for the regularization term in MMR, MR-Trace as well as λ of MMR, in the range of $\{10^{-5}, 10^{-4}, \ldots, 10^{5}\}$ and report the best result *w.r.t.* each method. We tune the hyper-parameters via 5-fold cross validation and report the best parameter *w.r.t.* RMSE for each method.

Comparison on Memory Impairment Prediction

We summarize the comparison results of cognitive score prediction in Tables 1 and 2, where we mark the best results in bold. In Table 1, we perform the t-test to compare the MMR result in each data with other methods to show if the advantage is significant *w.r.t.* the *p*-value. We set the significance level as 0.10. In each data, we mark the methods with significant difference with the “$*$” sign.

From these results we can notice that MMR performs equal or better than all the comparing methods, which indicates the advantage of adopting nonlinear regression in our model. We can find that in Fluency and RAVLT, MMR outperforms other methods with statistically significant advantage. In the TRAILS data, even though MMR gets slight higher RMSE value than MR-Trace, the difference is not statistically significant. It confirms that MMR is more suitable to describe the complicated relationship between imaging features and cognitive scores. In addition, the prediction results of MMR show the effectiveness of finding low-rank structure among multiple learning tasks. Since the number of tasks in the data is small (two tasks in Fluency while three tasks in RAVLT and TRAILS), the advantage of MMR may not be shown significantly. We can notice that MMR performs the best in the RAVLT data, as the MMR model is able to explore and utilize the interrelations among multiple learning tasks and improve the overall performance. In addition, we find that the range of the target variable in TRAILS data is much larger (the range in TRAILS data is 300 while the range for the other two data is smaller than 100). Since our model could better fit the task relationship among the training data, a larger testing data range may introduce higher testing error if the training data is slightly overfitted. The experimental results have validated the effectiveness of the MMR for simultaneously handling nonlinear relationship between neuroimaging biomarkers and cognitive scores and the correlation among the scores. The high performance of the proposed MMR and its huge advantages over previous representative regression models indicate its great potential to conduct even more challenging multivariate prediction tasks in clinical practice.

Stability of the MMR Method

In this subsection, we show the stability of the MMR method when we set the number of nodes in the hidden layer as different values. From the results in Fig. 2 we can find that the MMR results are quite stable *w.r.t.* different number of nodes. This is important in real applications since MMR does not require much effort in tuning the hyper-parameters.

Conclusion

In this paper, we have presented a new multi-target regression model, called multi-layer multi-target regression (MMR), for cognitive assessment of Alzheimer's disease. The MMR is able to simultaneously handle the nonlinear relationship between neuraoimaging biomarkers and cognitive assessment scores and the inter-correlation among the scores, which can largely improve the prediction performance. The MMR has been evaluated by extensive experiments on the public ADNI database, and produced high prediction performance surpassing most of the previous representative regression models. The results have shown the great effectiveness of the MMR in cognitive assessment prediction, which indicates its great potential for multi-target prediction in clinical prediction.

Information Sharing Statement

The data used in this paper can be obtained from the ADNI database (RRID:SCR 003007, adni.loni.usc.edu). An executable program of our model is available upon request.

Acknowledgements

This work was partially supported by the following grants: NSF-DBI 1356628, NSF-IIS 1633753, NIH R01 AG049371.

Appendix

Proof By the definition of the nuclear norm, we can re-write it in terms of traces as follows

$$
\begin{aligned}
\|S\|_* &= tr(\sqrt{S^\top S}) = tr(\sqrt{\left(U\sum V^\top\right)^\top\left(U\sum V^\top\right)}) \qquad (20)\\
&= tr(\sqrt{V\sum^\top U^\top U\sum V^\top} = tr\,(\sqrt{V\sum^\top \sum V^\top})\\
&= tr(\sqrt{V\sum^\top \sum V^\top})\\
&= tr(\sqrt{V\sum V^\top V\sum V^\top}\\
&= tr(V\sum V^\top)\\
&= tr\left(\sum\right)
\end{aligned}
$$

Therefore, the nuclear norm of *S* can be also defined as the sum of the singular value decomposition of *S*. From (13), we have

$$\partial S = \partial U \sum V^{\top} + U \partial \sum V^{\top} + U \sum \partial V^{\top}, \quad (21)$$

which gives rise to

$$U \partial \sum V^{\top} = \partial S - \partial U \sum V^{\top} - U \sum \partial V^{\top}. \quad (22)$$

Multiplying $U^{\top}$ on both sides of (22), we have

$$U^{\top} U \partial \sum V^{\top} V = U^{\top} \partial S V - U^{\top} \partial U \sum V^{\top} V - U^{\top} U \sum \partial V^{\top} V \quad (23)$$

Since *U* is also an orthogonal matrix, we achieve

$$\partial \sum = U^{\top} \partial S V - U^{\top} \partial U \sum - \sum \partial V^{\top} V. \quad (24)$$

Note that we have the fact that

$$0 = \partial I = \partial\left(U^{\top} U\right) = \partial U^{\top} U + U^{\top} \partial U, \quad (25)$$

where *I* is an identity matrix, and therefore $U^{\top} \partial U$ is an antisymmetric matrix. We have

$$\begin{aligned} tr\left(U^{\top} \partial U \sum\right) &= tr\left(\left(U^{\top} \partial U \sum\right)^{\top}\right) = tr\left(\sum^{\top} \partial U^{\top} U\right) \quad (26) \\ &= -tr\left(\sum U^{\top} \partial U\right) = -tr\left(U^{\top} \partial U \sum\right) \end{aligned}$$

which indicates that $tr(U^{\top} \partial U \Sigma) = 0$. Similarly, we also have $tr(\Sigma \partial V^{\top} V) = 0$. Therefore, we achieve

$$tr\left(\partial \sum\right) = tr\left(U^{\top} \partial S V\right) \quad (27)$$

By taking the derivative of $\|S\|_*$ w.r.t. *S*, we obtain

$$\frac{\partial \|S\|_*}{\partial S} = \frac{tr(\partial \sum)}{\partial S} = \frac{tr\left(U^{\top} \partial S V\right)}{\partial S} U V^{\top} \quad (28)$$

which closes the proof.

References

Agarwal A, Gerber S, Daume H (2010). Learning multiple tasks using manifold regularization. In Advances in neural information processing system (pp. 46–54).

Aho T, Ženko B, Džeroski S, Elomaa T (2012). Multi-target regression with rule ensembles. Journal of Machine Learning Research, 13(1), 2367–2407.

Alvarez M, Rosasco L, Lawrence N (2012). Kernels for vectorvalued functions: a review Foundations and Trends in Machine Learning.

Argyriou A, Evgeniou T, Pontil M (2008). Convex multi-task feature learning. Machine Learning, 73(3), 243–272.

Armijo L (1966). Minimization of functions having lipschitz continuous first partial derivatives. Pacific Journal of Mathematics, 16(1), 1–3.

Association A et al. (2016). 2016 alzheimer's disease facts and figures. Alzheimer's & Dementia, 12(4), 459–509.

Ciliberto C, Mroueh Y, Poggio T, Rosasco L (2015). Convex learning of multiple tasks and their structure. In Internationl conference on machine learning (pp. 1548–1557).

DauméIII H (2009). Bayesian multitask learning with latent hierarchies. In Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence (pp. 135–142).

Dinuzzo F (2013). Learning output kernels for multi-task problems. Neurocomputing, 118, 119–126.

Dinuzzo F, Ong CS, Pillonetto G, Gehler PV (2011). Learning output kernels with block coordinate descent. In Internationl conference on machine learning (pp. 49–56).

Dinuzzo F, & Schölkopf B (2012). The representer theorem for Hilbert spaces: a necessary and sufficient condition. In Advances in neural information processing system (pp. 189–196).

Evgeniou T, Micchelli CA, Pontil M (2005). Learning multiple tasks with kernel methods. In Journal of machine learning research (pp. 615–637).

Falahati F, Ferreira D, Muehlboeck JS, Eriksdotter M, Simmons A, Wahlund LO, Westman E (2016). Longitudinal investigation of an mri-based alzheimers disease diagnostic index in adni. Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 12(7), P732–P733.

Feng Y, Lv SG, Hang H, Suykens JA (2016). Kernelized elastic net regularization: generalization bounds, and sparse recovery. Neural Computation, 28(3), 525–562. [PubMed: 26735744]

Ferrarini L, Palm WM, Olofsen H, van der Landen R, Blauw GJ, Westendorp RG, Bollen EL, Middelkoop HA, Reiber JH, van Buchem MA, et al. (2008). Mmse scores correlate with local ventricular enlargement in the spectrum from cognitively normal to alzheimer disease. NeuroImage, 39(4), 1832–1838. [PubMed: 18160312]

Folstein MF, Folstein SE, McHugh PR (1975). A mini-mental state: a practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research, 12(3), 189–198. [PubMed: 1202204]

Gillberg J, Marttinen P, Pirinen M, Kangas AJ, Soininen P, Ali M, Havulinna AS, Järvelin MR, Ala-Korpela M, Kaski S (2016). Multiple output regression with latent noise. The Journal of Machine Learning Research, 17(1), 4170–4204.

Gong P, Zhou J, Fan W, Ye J (2014). Efficient multi-task feature learning with calibration. In Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining (pp. 761–770).

Hara K, & Chellappa R (2014). Growing regression forests by classification: applications to object pose estimation. In European conference on computer vision (pp. 552–567).

Jack CR, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ, L Whitwell J, Ward C, et al. (2008). The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging, 27(4), 685–691. [PubMed: 18302232]

Kabani NJ (1998). 3d anatomical atlas of the human brain. Neuroimage, 7, P–0717.

Kolar M, Lafferty J, Wasserman L (2011). Union support recovery in multi-task learning. Journal of Machine Learning Research, 12, 2415–2435.

Kumar A, & Daume H (2012). Learning task grouping and overlap in multi-task learning. In Internationl conference on machine learning (pp. 1383–1390).

Lee SI, Chatalbashev V, Vickrey D, Koller D (2007). Learning a meta-level prior for feature relevance from multiple related tasks. In Internationl conference on machine learning (pp. 489–496).

Li C, Georgiopoulos M, Anagnostopoulos GC (2015). Pareto-path multitask multiple kernel learning. IEEE Transactions on Neural Networks and Learning Systems, 26(1), 51–61. [PubMed: 25532155]

Li H, Chen N, Li L (2012). Error analysis for matrix elastic-net regularization algorithms. IEEE Transactions on Neural Networks and Learning Systems, 23(5), 737–748. [PubMed: 24806123]

Liu H, Wang L, Zhao T (2014). Multivariate regression with calibration. In Advances in neural information processing system (pp. 127–135).

Lounici K, Pontil M, Van De Geer S, Tsybakov AB (2011). Oracle inequalities and optimal inference under group sparsity. In The annals of statistics (pp. 2164–2204).

Molstad AJ, & Rothman AJ (2015). Indirect multivariate response linear regression. arXiv: 1507.04610.

Moradi E, Hallikainen I, Hänninen T, Tohka J, Initiative ADN, et al. (2016). Rey’s auditory verbal learning test scores can be predicted from whole brainmri in alzheimer’s disease NeuroImage: Clinical.

Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, Trojanowski JQ, Toga AW, Beckett L (2005). Ways toward an early diagnosis in alzheimers disease: the alzheimers disease neuroimaging initiative (adni). Alzheimer’s & Dementia, 1(1), 55–66.

Pan Y, Xia R, Yin J, Liu N (2015). A divide-and-conquer method for scalable robust mul-titask learning. IEEE Transactions on Neural Networks and Learning Systems, 26(12), 3163–3175. [PubMed: 25775500]

Rai P, Kumar A, Daume H (2012). Simultaneously leveraging output and task structures for multiple-output regression. In Advances in neural information processing system (pp. 3185–3193).

Rakitsch B, Lippert C, Borgwardt K, Stegle O (2013). It is all in the noise: efficient multi-task gaussian process inference with structured residuals. In NIPS (pp. 1466–1474).

Rothman AJ, Levina E, Zhu J (2010). Sparse multivariate regression with covariance estimation. Journal of Computational and Graphical Statistics, 19(4), 947–962. [PubMed: 24963268]

Schmidt M, et al. (1996). Rey auditory verbal learning test: a handbook Western Psychological Services Los Angeles.

Seshadri S, DeStefano AL, Au R, Massaro JM, Beiser AS, Kelly-Hayes M, Kase CS, D’Agostino RB, DeCarli C, Atwood LD, et al. (2007). Genetic correlates of brain aging on mri and cognitive test measures: a genome-wide association and linkage analysis in the framingham study. BMC Medical Genetics, 8(1), S15. [PubMed: 17903297]

Shen D, & Davatzikos C (2002). Hammer: hierarchical attribute matching mechanism for elastic registration. IEEE Transactions on Medical Imaging, 21(11), 1421–1439. [PubMed: 12575879]

Sled JG, Zijdenbos AP, Evans AC (1998). A nonparametric method for automatic correction of intensity nonuniformity in mri data. IEEE Transactions on Medical Imaging, 17(1), 87–97. [PubMed: 9617910]

Sohn KA, & Kim S (2012). Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In International conference on artificial intelligence and statistics (pp. 1081–1089).

Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pp. 267–288.

Tsoumakas G, Spyromitros-Xioufis E, Vrekou A, Vlahavas I (2014). Multi-target regression via random linear target combinations. In Machine learning and knowledge discovery in databases (pp. 225–240). Springer.

Wang H, Nie F, Huang H, Yan J, Kim S, Risacher S, Saykin A, Shen L (2012). High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer’s disease progression prediction. In Advances in neural information processing systems (pp. 1277–1285).

Wang Y, Nie J, Yap PT, Li G, Shi F, Geng X, Guo L, Shen D, Initiative ADN, et al. (2014). Knowledge-guided robust mri brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates. PloS One, 9(1), e77810. [PubMed: 24489639]

Wang Y, Nie J, Yap PT, Shi F, Guo L, Shen D (2011). Robust deformable-surface-based skull-stripping for large-scale studies. In Medical image computing and computer-assisted intervention– MICCAI 2011 (pp. 635–642). Springer.

Yu K, Tresp V, Schwaighofer A (2005). Learning gaussian processes from multiple tasks. In International conference on machine learning (pp. 1012–1019).

Zhang Y, Brady M, Smith S (2001). Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. IEEE Transactions on Medical Imaging, 20(1), 45–57. [PubMed: 11293691]

Zhang Y, & Yeung DY (2013). Learning high-order task relationships in multi-task learning. In International joint conference on artificial intelligence (pp. 1917–1923).

Zhang Y, & Yeung DY (2014). A regularization approach to learning task relationships in multitask learning. ACM Transactions on Knowledge Discovery from Data, 8(3), 12.

Zhen X, Yu M, He X, Li S (2017). Multi-target regression via robust low-rank learning. In IEEE transactions on pattern analysis and machine Intelligence

Zhou Q, & Zhao Q (2016). Flexible clustered multi-task learning by learning representative tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2), 266–278. [PubMed: 26761733]

Zou H, & Hastie T (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320.

Zhu X, Li X, Zhang S, Ju C, Wu X (2017). Robust joint graph sparse coding for unsupervised spectral feature selection. IEEE Transactions on Neural Networks and Learning systems, 28(6), 1263–1275. [PubMed: 26955053]

Zhu X, Suk HI, Wang L, Lee SW, Shen D (2015). Alzheimer’s disease neuroimaging initiative: a novel relational regularization feature selection method for joint regression and classification in AD diagnosis. Medical Image Analysis, 38, 205–214. [PubMed: 26674971]

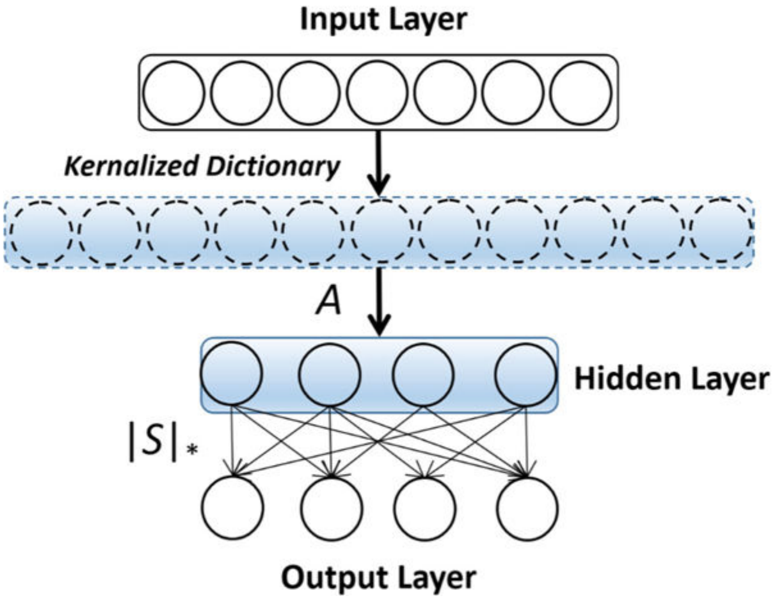

Fig. 1.

The learning architecture of the multi-layer multi-target regression (MMR) model

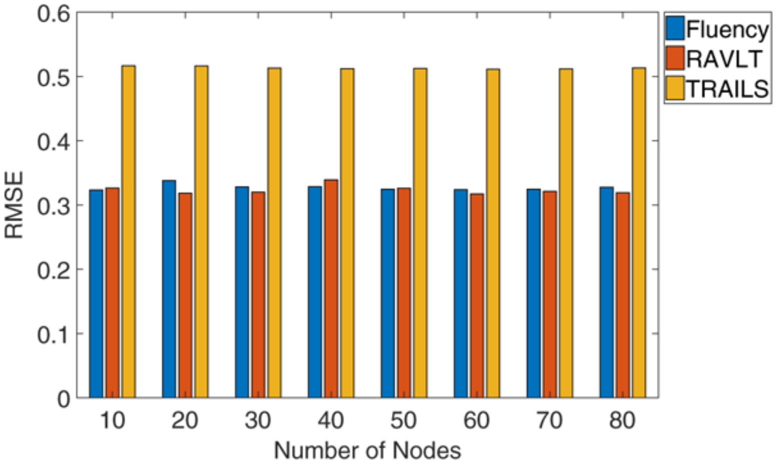

Fig. 2.

RMSE results with different number of hidden nodes in the MMR model

Table 1

RMSE results of cognitive score prediction using MRI biomarkers. The “*” sign indicates statistically significant difference between MMR result and other methods. The significance level is 0.10

Methods	Fluency	RAVLT	TRAILS
LSR	0.3395 ± 0.0159	0.3304 ± 0.0182	0.5284± 0.0273
MRR	0.3261 ± 0.0165(*)	0.3192 ± 0.0170(*)	0.5007± 0.0242
MR-Trace	0.3251 ± 0.0158	0.3191 ± 0.0165(*)	0.5008± 0.0303
MMR	0.3235± 0.0163	0.3172 ± 0.0153	0.5113± 0.0153

Table 2

CorCoe results of cognitive score prediction using MRI biomarkers

Methods	Fluency	RAVLT	TRAILS
LSR	0.5202 ± 0.0243	0.8776± 0.0129	0.5658±0.0556
MRR	0.5451 ± 0.0207	0.8851± 0.0115	0.6030±0.0425
MR-Trace	0.5485 ± 0.0185	0.8852± 0.0112	0.6045±0.0542
MMR	0.5427 ± 0.0153	0.8865± 0.0101	0.5796±0.0311